Journal of Physics Communications

PAPER

Augmented NETT regularization of inverse problems

Daniel Obmann[1], **Linh Nguyen**[2], **Johannes Schwab**[1] and **Markus Haltmeier**[1,*]

[1] Department of Mathematics, University of Innsbruck, Technikerstrasse 13, 6020 Innsbruck, Austria
[2] Department of Mathematics, University of Idaho, Moscow, ID 83844, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: daniel.obmann@uibk.ac.at, lnguyen@uidaho.edu, schwab@mrc-lmb.cam.ac.uk and markus.haltmeier@uibk.ac.at

Keywords: inverse problems, learned regularizer, computed tomography, neural networks, regularization

OPEN ACCESS

RECEIVED
16 June 2021

REVISED
6 September 2021

ACCEPTED FOR PUBLICATION
14 September 2021

PUBLISHED
4 October 2021

Abstract

We propose aNETT (augmented NETwork Tikhonov) regularization as a novel data-driven reconstruction framework for solving inverse problems. An encoder-decoder type network defines a regularizer consisting of a penalty term that enforces regularity in the encoder domain, augmented by a penalty that penalizes the distance to the signal manifold. We present a rigorous convergence analysis including stability estimates and convergence rates. For that purpose, we prove the coercivity of the regularizer used without requiring explicit coercivity assumptions for the networks involved. We propose a possible realization together with a network architecture and a modular training strategy. Applications to sparse-view and low-dose CT show that aNETT achieves results comparable to state-of-the-art deep-learning-based reconstruction methods. Unlike learned iterative methods, aNETT does not require repeated application of the forward and adjoint models during training, which enables the use of aNETT for inverse problems with numerically expensive forward models. Furthermore, we show that aNETT trained on coarsely sampled data can leverage an increased sampling rate without the need for retraining.

1. Introduction

Various applications in medical imaging, remote sensing and elsewhere require solving inverse problems of the form

$$y^\delta = \mathbf{K}x + \eta^\delta, \tag{1.1}$$

where $\mathbf{K}\colon \mathbb{X} \to \mathbb{Y}$ is an operator between Hilbert spaces modeling the forward problem, η^δ is the data perturbation, $y^\delta \in \mathbb{Y}$ is the noisy data and $x \in \mathbb{X}$ is the sought for signal. Inverse problems are well analyzed and several established approaches for its stable solution exist [1, 2]. Recently, neural networks and deep learning appeared as new paradigms for solving inverse problems [3–7]. Several approaches based on deep learning have been developed, including post-processing networks [8–12], regularizing null-space networks [13, 14], plug-and-play priors [15–17], deep image priors [18, 19], variational networks [20, 21], network cascades [22, 23], learned iterative schemes [24–29] and learned regularizers [30–33].

Classical deep learning approaches may lack data consistency for unknowns very different from the training data. To address this issue, in [31] a deep learning approach named NETT (NETwork Tikhonov) regularization has been introduced which considers minimizers of the NETT functional

$$\mathcal{N}_{\alpha,y^\delta}(x) := \mathcal{D}(\mathbf{K}x, y^\delta) + \alpha\mathcal{Q}(\mathbf{E}(x)). \tag{1.2}$$

Here, $\mathcal{D}$ is a similarity measure, $\mathbf{E}\colon \mathbb{X} \to \Xi$ is a trained neural network, $\mathcal{Q}\colon \Xi \to [0, \infty]$ a functional and $\alpha > 0$ a regularization parameter. In [31] it is shown that under suitable assumptions, NETT yields a convergent regularization method. This in particular includes provable stability guarantees and error estimates. Moreover, a training strategy has been proposed, where $\mathbf{E}$ is trained such that $\mathcal{Q}\circ\mathbf{E}$ favors artifact-free reconstructions over reconstructions with artifacts.

1.1. The augmented NETT

One of the main assumptions for the analysis of [31] is the coercivity of the regularizer $\mathcal{Q}\circ\mathbf{E}$ which requires special care in network design and training. In order to overcome this limitation, we propose an augmented form of the regularizer for which we are able to rigorously prove coercivity. More precisely, for fixed $c > 0$, we consider minimizers x_α^δ of the augmented NETT functional

$$\mathcal{A}_{\alpha,y^\delta}(x) := \mathcal{D}(\mathbf{K}x, y^\delta) + \alpha\left(\mathcal{Q}(\mathbf{E}(x)) + \frac{c}{2}\,\|x - (\mathbf{D}\circ\mathbf{E})(x)\|_2^2\right). \tag{1.3}$$

Here, $\mathcal{D}\colon \mathbb{Y}\times\mathbb{Y}\to[0,\infty]$ is a similarity measure and $\mathbf{D}\circ\mathbf{E}\colon \mathbb{X}\to\mathbb{X}$ is an encoder-decoder network trained such that for any signal x on a signal manifold we have $(\mathbf{D}\circ\mathbf{E})(x) \simeq x$ and that $\mathcal{Q}(\mathbf{E}(x))$ is small. We term this approach augmented NETT (aNETT) regularization. In this work we provide a mathematical convergence analysis for aNETT, present a novel modular training strategy and investigate its practical performance.

The term $\mathcal{Q}(\mathbf{E}(x))$ implements learned prior knowledge on the encoder coefficients, while smallness of $\|x - (\mathbf{D}\circ\mathbf{E})(x)\|_2^2$ forces x to be close to the signal manifold. The latter term also guarantees the coercivity of (1.3). In the original NETT version (1.2), coercivity of the regularizer requires coercivity conditions on the network involved. Indeed, in the numerical experiments, the authors of [31] observed a semi-convergence behaviour when minimizing (1.2), so early stopping of the iterative minimization scheme has been used as additional regularization. We attribute this semi-convergence behavior to a potential non-coercivity of the regularization term. In the present paper we address this issue systematically by augmentation of the NETT functional which guarantees coercivity and allows a more stable minimization. Coercivity is also one main ingredient for the mathematical convergence analysis.

An interesting practical instance of aNETT takes $\mathcal{Q}$ as a weighted ℓ^q-norm enforcing sparsity of the encoding coefficients [34, 35]. An important example for the similarity measure is given by the squared norm distance, which from a statistical viewpoint can be motivated by a Gaussian white noise model. General similarity measures allow us to adapt to different noise models which can be more appropriate for certain problems.

1.2. Main contributions

The contributions of this paper are threefold. As described in more detail below, we introduce the aNETT framework, mathematically analyze its convergence, and propose a practical implementation that is applied to tomographic limited data problems.

- The first contribution is to introduce the structure of the aNETT regularizer $\mathcal{R}(x) = \mathcal{Q}(\mathbf{E}(x)) + (c/2)\,\|x - (\mathbf{D}\circ\mathbf{E})(x)\|_2^2$. A similar approach has been studied in [36] for a linear encoder $\mathbf{E}$. However, in this paper we do not assume that the image x consists of two components u and v but rather assume that there is some transformation $\mathbf{E}$ in which the signal x has some desired property such as, for example, sparsity. The term $\mathcal{Q}(\mathbf{E}(x))$ enforces regularity of the analysis coefficients, which is an ingredient in most of existing variational regularization techniques. For example, this includes sparse regularization in frames or dictionaries, regularization with Sobolev norms or total variation regularization. On the other hand, the augmented term $\|x - (\mathbf{D}\circ\mathbf{E})(x)\|_2^2$ penalized distance to the signal manifold. It is the combination of these two terms that results in a stable reconstruction scheme without the need of strong assumptions on the involved networks.
- The second main contribution is the theoretical analysis of aNETT (1.3) in the context of regularization theory. We investigate the case where the image domain of the encoder is given by $\Xi = \ell^2(\Lambda)$ for some countable set Λ, and $\mathcal{Q}$ is a coercive functional measuring the complexity of the encoder coefficients. The presented analysis is in the spirit of the analysis of NETT given in [31]. However, opposed to NETT, the required coercivity property is derived naturally for the class of considered regularizers. This supports the use of the regularizer $\mathcal{R}$ also from a theoretical side. Moreover, the convergence rates results presented here use assumptions significantly different from [31]. While we present our analysis for the transform domain $\Xi = \ell^2(\Lambda)$ we could replace the encoder space by a general Hilbert or Banach space.
- As a third main contribution we propose a modular strategy for training $\mathbf{D}\circ\mathbf{E}$ together with a possible network architecture. First, independent of the given inverse problem, we train a $\mathcal{Q}$-penalized autoencoder that learns representing signals from the training data with low complexity. In the second step, we train a task-specific network which can be adapted to the specific inverse problem at hand. In our numerical experiments, we empirically found this modular training strategy to be superior to directly adapting the autoencoder to the inverse problem. For the $\mathcal{Q}$-penalized autoencoder, we train the modified version described in [37] of the tight frame U-Net of [38] in a way such that $\mathcal{Q}$ poses additional constraints on the autoencoder during the training process.

1.3. Outline

In section 2 we present the mathematical convergence analysis of aNETT. In particular, as an auxiliary result, we establish the coercivity of the regularization term. Moreover, we prove stability and derive convergence rates. Section 3 presents practical aspects for aNETT. We propose a possible architecture and training strategy for the networks, and a possible ADMM based scheme to obtain minimizers of the aNETT functional. In section 4, we present reconstruction results and compare aNETT with other deep learning based reconstruction methods. The paper concludes with a short summary and discussion. Parts of this paper were presented at the ISBI 2020 conference and the corresponding proceedings [39]. Opposed to the proceedings, this article treats a general similarity measure $\mathcal{D}$ and considers a general complexity measure $\mathcal{Q}$. Further, all proofs and all numerical results presented in this paper are new.

2. Mathematical analysis

In this section we prove the stability and convergence of aNETT as regularization method. Moreover, we derive convergence rates in the form of quantitative error estimates between exact solutions for noise-free data and aNETT regularized solutions for noisy data. To this end we make the assumption that we can achieve global minimizers of the functional (1.3) and analyze the properties of these solutions. This is a common assumption in variational regularization approaches and is adopted in this work. Extending the analysis to consider only local minima is consiberably more difficult and is out of the scope of this paper.

2.1. Assumptions and coercivity results

For our convergence analysis we make use of the following assumptions on the underlying spaces and operators involved.

Condition 2.1.

(A1) $\mathbb{X}$ and $\mathbb{Y}$ are Hilbert spaces.

(A2) $\Xi = \ell^2(\Lambda)$ for a countable Λ.

(A3) $\mathbf{K}$: $\mathbb{X} \to \mathbb{Y}$ is weakly sequentially continuous.

(A4) $\mathbf{E}$: $\mathbb{X} \to \Xi$ is weakly sequentially continuous.

(A5) $\mathbf{D}$: $\Xi \to \mathbb{X}$ is weakly sequentially continuous.

(A6) $\mathcal{Q}$: $\Xi \to [0, \infty]$ is coercive and weakly sequentially lower semi-continuous.We set $\mathbf{N} := \mathbf{D}\circ\mathbf{E}$ and, for given $c > 0$, define

$$\mathcal{R}\colon \mathbb{X} \to [0, \infty]\colon x \mapsto \mathcal{Q}(\mathbf{E}(x)) + \frac{c}{2}\,\|x - \mathbf{N}(x)\|_2^2, \tag{2.1}$$

which we refer to as the aNETT (or augmented NETT) regularizer.

According to (A4)–(A6), the aNETT regularizer is weakly sequentially lower semi-continuous. As a main ingredient for our analysis we next prove its coercivity.

Theorem 2.2 Coercivity of the aNETT regularizer. *If Condition 2.1 holds, then the regularizer* $\mathcal{R}\colon \mathbb{X} \to [0, \infty]$ *as defined in (2.1) with* $\mathbf{N} = \mathbf{D}\circ\mathbf{E}$ *is coercive.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be some sequence in $\mathbb{X}$ such that $(\mathcal{R}(x_n))_{n\in\mathbb{N}}$ is bounded. Then by definition of $\mathcal{R}$ it follows that $(\mathcal{Q}(\mathbf{E}(x_n)))_{n\in\mathbb{N}}$ is bounded and by coercivity of $\mathcal{Q}$ we have that $(\mathbf{E}(x_n))_{n\in\mathbb{N}}$ is also bounded. By assumption, $\mathbf{D}$ is weakly sequentially continuous and thus $(\|\mathbf{N}(x_n)\|)_{n\in\mathbb{N}}$ must be bounded. Using that $c > 0$, we obtain the inequality $\|x_n\|^2 \leqslant 2\,\|x_n - \mathbf{N}(x_n)\|^2 \;+\; 2\,\|\mathbf{N}(x_n)\|^2 \leqslant (4/c)\,\mathcal{R}(x_n) + 2\,\|\mathbf{N}(x_n)\|^2$. This shows that $(x_n)_{n\in\mathbb{N}}$ is bounded and therefore that $\mathcal{R}$ is coercive. □

Example 2.3 Sparse aNETT regularizer. To obtain a sparsity promoting regularizer we can choose $\mathcal{Q}(\xi) = \|\xi\|_{1;w} := \sum_{\lambda\in\Lambda} w_\lambda\,|\xi_\lambda|^q$ where $q \in [1, 2]$ and $\inf_\lambda w_\lambda > 0$. Since $q \in [1, 2]$ we have $\|\cdot\|^2 \leqslant (\inf_\lambda w_\lambda)^{-1/q}\;\|\cdot\|_{1;w}$ and hence $\|\cdot\|_{1;w}$ is coercive. As a sum of weakly sequentially lower semi-continuous functionals it is also weakly sequentially lower semi-continuous [35]. Therefore Condition (A6) is satisfied for the weighted ℓ^q-norm. Together with theorem 2.2, we conclude that the resulting weighted sparse aNETT regularizer $x \mapsto \|\mathbf{E}(x)\|_{1;w} + (c/2)\,\|x - (\mathbf{D}\circ\mathbf{E})(x)\|_2^2$ is a coercive and weakly sequentially lower semi-continuous functional.

For the further analysis we will make the following assumptions regarding the similarity measure $\mathcal{D}$: $\mathbb{Y} \times \mathbb{Y} \to [0, \infty]$.

Condition 2.4 Similarity measure.

(B1) $\forall y_0, y_1$: $\mathcal{D}(y_0, y_1) = 0 \Leftrightarrow y_0 = y_1$.

(B2) $\mathcal{D}$ is sequentially lower semi-continuous with respect to the weak topology in the first and the norm topology in the second argument.

(B3) $\forall (y_n)_{n\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$: $(\mathcal{D}(y, y_n) \to 0 \Rightarrow y_n \to y \text{ as } n \to \infty)$.

(B4) $\mathcal{D}(y, y_n) \to 0$ as $n \to \infty \Rightarrow (\forall z \in \mathbb{Y}$: $\mathcal{D}(z, y) < \infty \Rightarrow \mathcal{D}(z, y_n) \to \mathcal{D}(z, y))$.

(B5) $\forall y \in \mathbb{Y}\ \forall \alpha > 0$: $\exists x \in \mathbb{X}$ with $\mathcal{D}(\mathbf{K}x, y) + \alpha\mathcal{R}(x) < \infty$.

While (B1)–(B4) restrict the choice of the similarity measure, (B5) is a technical assumption involving the forward operator, the regularizer and the similarity measure, that is required for the existence of minimizers. For a more detailed discussion of these assumptions we refer to [40].

Example 2.5 Similarity measures using the norm. A classical example of a similarity measure satisfying (B1)–(B4) is given by $\mathcal{D}(y_0, y_1) = \|y_0 - y_1\|^p$ for some $p \geqslant 1$ and more generally by $\mathcal{D}(y_0, y_1) = \psi(\|y_0 - y_1\|)$, where ψ: $[0, \infty) \to [0, \infty)$ is a continuous and monotonically increasing function that satisfies $\forall t \geqslant 0$: $\psi(t) = 0 \Leftrightarrow t = 0$.

Taking into account theorem 2.2, Conditions 2.1 and 2.4 imply that the aNETT functional $\mathcal{A}_{\alpha,y}$ defined by (1.3), (2.1) is proper, coercive and weakly sequentially lower semi-continuous. This in particular implies the existence of minimizers of $\mathcal{A}_{\alpha,y}$ for all data $y \in \mathbb{Y}$ and regularization parameters $\alpha > 0$ (compare) [2, 31].

2.2. Stability

Next we prove the stability of minimizing the aNETT functional $\mathcal{A}_{\alpha,y}$ regarding perturbations of the data y.

Theorem 2.6 Stability. *Let Conditions 2.1 and 2.4 hold, $y \in \mathbb{Y}$ and $\alpha > 0$. Moreover, let $(y_n)_{n\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$ be a sequence of perturbed data with $\mathcal{D}(y, y_n) \to 0$ and consider minimizers $x_n \in \arg\min \mathcal{A}_{\alpha,y_n}$. Then the sequence $(x_n)_{n\in\mathbb{N}} \in \mathbb{X}^{\mathbb{N}}$ has at least one weak accumulation point and weak accumulation points are minimizers of $\mathcal{A}_{\alpha,y}$. Moreover, for any weak accumulation point $x^\ddagger$ of $(x_n)_{n\in\mathbb{N}}$ and any subsequence $(x_{\tau(n)})_{n\in\mathbb{N}}$ with $x_{\tau(n)} \rightharpoonup x^\ddagger$ we have $\mathcal{R}(x_{\tau(n)}) \to \mathcal{R}(x^\ddagger)$.*

Proof. Let $x_\star \in \mathbb{X}$ be such that $\mathcal{A}_{\alpha,y}(x_\star) < \infty$. By definition of x_n we have $\mathcal{A}_{\alpha,y_n}(x_n) \leqslant \mathcal{A}_{\alpha,y_n}(x_\star)$. Since by assumption $\mathcal{D}(y, y_n) \to 0$ and $\mathcal{D}(\mathbf{K}x_\star, y) < \infty$, we have $\mathcal{D}(\mathbf{K}x_\star, y_n) \to \mathcal{D}(\mathbf{K}x_\star, y)$. This implies that $\mathcal{A}_{\alpha,y_n}(x_\star)$ is bounded by some positive constant $m_\star$ for sufficiently large n. By definition of $\mathcal{A}_{\alpha,y_n}$ we have $\alpha\mathcal{R}(x_n) \leqslant \mathcal{A}_{\alpha,y_n}(x_n) \leqslant m_\star + \alpha\mathcal{R}(x_\star)$. Since $\mathcal{R}$ is coercive it follows that $(x_n)_{n\in\mathbb{N}}$ is a bounded sequence and hence it has a weakly convergent subsequence.

Let $(x_{\tau(n)})_{n\in\mathbb{N}}$ be a weakly convergent subsequence of $(x_n)_{n\in\mathbb{N}}$ and denote its limit by $x^\ddagger$. By the lower semi-continuity we get $\mathcal{D}(\mathbf{K}x^\ddagger, y) \leqslant \liminf_{n\to\infty} \mathcal{D}(\mathbf{K}x_{\tau(n)}, y_{\tau(n)})$ and $\mathcal{R}(x^\ddagger) \leqslant \liminf_{n\to\infty} \mathcal{R}(x_{\tau(n)})$. Thus for all $x \in \mathbb{X}$ with $\mathcal{A}_{\alpha,y}(x) < \infty$ we have

$$\begin{aligned}&\mathcal{A}_{\alpha,y}(x^\ddagger) \leqslant \liminf_{n\to\infty} \mathcal{D}(\mathbf{K}x_{\tau(n)}, y_{\tau(n)}) + \alpha \liminf_{n\to\infty} \mathcal{R}(x_{\tau(n)})\\ &\leqslant \limsup_{n\to\infty} \mathcal{D}(\mathbf{K}x_{\tau(n)}, y_{\tau(n)}) + \alpha\mathcal{R}(x_{\tau(n)}) \leqslant \limsup_{n\to\infty} \mathcal{A}_{\alpha,y_{\tau(n)}}(x) = \mathcal{A}_{\alpha,y}(x).\end{aligned}$$

This shows that $x^\ddagger \in \arg\min \mathcal{A}_{\alpha,y}$ and, by considering $x = x^\ddagger$ in the above displayed equation, that $\mathcal{A}_{\alpha,y_{\tau(n)}}(x_{\tau(n)}) \to \mathcal{A}_{\alpha,y}(x^\ddagger)$. Moreover, we have

$$\begin{aligned}&\limsup_{n\to\infty} \alpha\mathcal{R}(x_{\tau(n)}) \leqslant \limsup_{n\to\infty} \mathcal{A}_{\alpha,y_{\tau(n)}}(x_{\tau(n)}) - \liminf_{n\to\infty} \mathcal{D}(\mathbf{K}x_{\tau(n)}, y_{\tau(n)})\\ &\leqslant \mathcal{A}_{\alpha,y}(x^\ddagger) - \mathcal{D}(\mathbf{K}x^\ddagger, y) = \alpha\mathcal{R}(x^\ddagger).\end{aligned}$$

This shows $\mathcal{R}(x_{\tau(n)}) \to \mathcal{R}(x^\ddagger)$ as $n \to \infty$ and concludes the proof. □

In the following we say that the similarity measure $\mathcal{D}$ satisfies the quasi triangle-inequality if there is some $q \geqslant 1$ such that

$$\forall y_0, y_1, y_2 \in \mathbb{Y}: \quad \mathcal{D}(y_0, y_1) \leqslant q \cdot (\mathcal{D}(y_0, y_2) + \mathcal{D}(y_2, y_1)). \tag{2.2}$$

While this property is essential for deriving convergence rate results, we will show below that it is not enough to guarantee stability of minimizing the augmented NETT functional in the sense of theorem 2.6. Note that [31] assumes the quasi triangle-inequality (2.2) instead of Condition (B4). The following remarks shows that (2.2) is not sufficient for the stability result of theorem 2.6 to hold and therefore Condition (B4) has to be added to the list of assumptions in [31] required for the stability.

Example 2.7 Instability in the absence of Condition (B4). Consider the similarity measure $\mathcal{D}: \mathbb{Y} \times \mathbb{Y} \to [0, \infty]$ defined by

$$\mathcal{D}(y_0, y_1) := H(y_1) \|y_0 - y_1\|^2, \tag{2.3}$$

where $H: \mathbb{Y} \to [0, 1]$ is defined by $H(y_1) = 1$ if $\|y_1\| \leqslant 1$ and $H(y_1) = 2$ otherwise. Moreover, choose $\mathbb{X} = \mathbb{Y}$, let $\mathbf{K} = \mathrm{Id}$ be the identity operator and suppose the regularizer takes the form $\mathcal{R} = \|\cdot\|^2$.

- The similarity measure defined in (2.3) satisfies (B1)–(B3): Convergence with respect to $\mathcal{D}$ is equivalent to convergence in norm which implies that (B3) is satisfied. Moreover, we have $\forall y_0, y_1: \mathcal{D}(y_0, y_1) = 0 \Leftrightarrow y_0 = y_1$, which is (B1). Consider sequences $z_n \to z$ and $y_n \rightharpoonup y$. The sequential lower semi-continuity stated in (B2) can be derived by separately looking at the cases $\|z\| \leqslant 1$ and $\|z\| > 1$. In the first case, by the continuity of the norm we have $\forall n \in \mathbb{N}: H(z) \leqslant H(z_n)$. In the second case, we have $\|y_n\| > 1$ for n sufficiently large. In both cases, the lower semi-continuity property follows from the weak lower semi-continuity property of the norm.
- The similarity measure defined by (2.3) does not satisfy (B4): To see this, we define $y_n := (1 + \delta_n)y$ where $\|y\| = 1$ and $\delta_n > 0$ is taken as a non-increasing sequence converging to zero. We have $y_n \to y$ and hence also $\mathcal{D}(y, y_n) \to 0$ as $n \to \infty$. For any $z \in \mathbb{Y}$ we have $\mathcal{D}(z, y) < \infty$ and $\mathcal{D}(z, y_n) = 2\|z - y_n\|^2 \to 2\|z - y\|^2 = 2\mathcal{D}(z, y)$ as $n \to \infty$. In particular, $\mathcal{D}(z, y_n)$ does not converge to $\mathcal{D}(z, y)$ if $z \neq y$ and therefore (B5) does not hold. In summary, all requirement for theorem 2.6 are satisfied, except of the continuity assumption (B4).
- We have $\|y_0 - y_1\|^2 \leqslant \mathcal{D}(y_0, y_1) \leqslant 2\|y_0 - y_1\|^2$ which implies that the similarity measure satisfies the quasi triangle-inequality (2.2). However as shown next, this is not sufficient for stable reconstruction in the sense of theorem 2.6. To that end, let $y_n := (1 + \delta_n)y$ with $\|y\| = 1$ and $\delta_n \downarrow 0$ and let $\alpha > 0$. In particular, $\mathcal{D}(x, y_n) = 2\|x - y_n\|^2$ and $\mathcal{D}(x, y) = \|x - y\|^2$. Therefore the minimizer of $\arg\min \mathcal{D}(\mathrm{x}, \mathrm{y}_n) + \alpha\|\mathrm{x}\|^2$ with perturbed data y_n is given by $x_n = y_n/(1 + \alpha/2)$ and the minimizer of $\mathcal{D}(x, y_n) + \alpha\|x\|^2$ for data y is given by $x^\ddagger = y/(1 + \alpha)$. We see that $x_n \to y/(1 + \alpha/2)$, which is clearly different from $x^\ddagger$. In particular, minimizing $\mathcal{D}(\cdot, y) + \alpha\|\cdot\|^2$ does not stably depend on data y. Theorem 2.6 states that stability holds if (B4) is satisfied.

While the above example may seem somehow constructed, it shows that one has to be careful when choosing the similarity measure in order to obtain a stable reconstruction scheme.

2.3. Convergence

In this subsection we consider the limit process as the noise-level δ tends to 0. Assuming that $y \in \mathrm{ran}(\mathbf{K})$ we would expect the regularized solutions to converge to some solution of the equation $\mathbf{K}x = y$. This raises the obvious question whether this solution has any additional properties. In fact, we prove that the minimizers of the aNETT functional for noisy data converge to such a special kind of solution, namely solutions which minimize $\mathcal{R}$ among all possible solutions. For that purpose, here and below we use the following notation.

Definiton 2.8 $\mathcal{R}$-minimizing solutions. For $y \in \mathbb{Y}$, we call an element $x^\ddagger \in \mathrm{dom}(\mathcal{R})$ an $\mathcal{R}$-minimizing solution of the equation $\mathbf{K}x = y$ if

$$x^\ddagger \in \arg\min\{\mathcal{R}(\mathrm{x}): \mathrm{x} \in \mathrm{dom}(\mathcal{R}) \wedge \mathbf{K}\mathrm{x} = \mathrm{y}\}.$$

An $\mathcal{R}$-minimizing solution always exists provided that data satisfies $y \in \mathbf{K}(\mathrm{dom}(\mathcal{R}))$, which means that the equation $\mathbf{K}x = y$ has at least one solution with finite value of $\mathcal{R}$. To see this, consider a sequence of solutions $(x_n)_{n\in\mathbb{N}}$ with $\mathcal{R}(x_n) \to \inf\{\mathcal{R}(x): \mathbf{K}x = y \wedge \mathcal{R}(x) < \infty\}$. Since $\mathcal{R}$ is coercive there exists a weakly convergent subsequence $(x_{\tau(n)})_{n\in\mathbb{N}}$ with weak limit $x^\ddagger$. Using the weak sequential lower semi-continuity of $\mathcal{R}$ one concludes that $x^\ddagger$ is an $\mathcal{R}$-minimizing solution. We first show weak convergence.

Theorem 2.9 Weak convergence of aNETT. *Suppose Conditions 2.1 and 2.4 are satisfied. Let* $y \in \mathbf{K}(\mathrm{dom}(\mathcal{R}))$, $(\delta_n)_{n\in\mathbb{N}} \in (0, \infty)^{\mathbb{N}}$ *with* $\delta_n \to 0$ *and let* $(y_n)_{n\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$ *satisfy* $\mathcal{D}(y, y_n) \leqslant \delta_n$. *Choose* $\alpha_n > 0$ *such that* $\lim_{n\to\infty} \delta_n/\alpha_n = \lim_{n\to\infty} \alpha_n = 0$ *and let* $x_n \in \arg\min \mathcal{A}_{\alpha_n, y_n}$. *Then the following hold:*

(a) $(x_n)_{n\in\mathbb{N}}$ *has at least one weakly convergent subsequence.*

(b) All accumulation points of $(x_n)_{n\in\mathbb{N}}$ *are* $\mathcal{R}$*-minimizing solutions of* $\mathbf{K}x = y$.

(c) For every convergent subsequence $(x_{\tau(n)})_{n\in\mathbb{N}}$ *it holds* $\mathcal{R}(x_{\tau(n)}) \to \mathcal{R}(x^\ddagger)$.

(d) If the $\mathcal{R}$*-minimizing solution* $x^\ddagger$ *is unique then* $x_n \rightharpoonup x^\ddagger$.

Proof. (a): Because $y \in \mathbf{K}(\mathrm{dom}(\mathcal{R}))$, there exists an $\mathcal{R}$-minimizing solution of the equation $\mathbf{K}x = y$ which we denote by $x^\ddagger$. Because $x_n \in \arg\min \mathcal{A}_{\alpha_n, y_n}$ we have

$$\begin{aligned}&\alpha_n \mathcal{R}(x_n) \leqslant \mathcal{D}(\mathbf{K}x_n, y_n) + \alpha_n \mathcal{R}(x_n) \leqslant \mathcal{D}(\mathbf{K}x^\ddagger, y_n) + \alpha_n \mathcal{R}(x^\ddagger)\\ &= \mathcal{D}(y, y_n) + \alpha_n \mathcal{R}(x^\ddagger) \leqslant \delta_n + \alpha_n \mathcal{R}(x^\ddagger).\end{aligned} \tag{2.4}$$

Because $\alpha_n, \delta_n \to 0$ this shows that $(\mathcal{R}(x_n))_{n\in\mathbb{N}}$ is bounded. Due to the coercivity of the aNETT regularizer (see theorem 2.2), this implies that $(x_n)_{n\in\mathbb{N}}$ has a weakly convergent subsequence.

(b), (c): Let $(x_{\tau(n)})_{n\in\mathbb{N}}$ be a weakly convergent subsequence of $(x_n)_{n\in\mathbb{N}}$ with limit $x_\star$. From the weak lower semi-continuity we get $\mathcal{D}(\mathbf{K}x_\star, y) \leqslant \liminf_{n\to\infty} \mathcal{D}(\mathbf{K}x_{\tau(n)}, y_{\tau(n)}) + \alpha_{\tau(n)} \mathcal{R}(x_{\tau(n)}) \leqslant \liminf_{n\to\infty} \mathcal{D}(\mathbf{K}x^\ddagger, y_{\tau(n)}) + \alpha_{\tau(n)} \mathcal{R}(x^\ddagger) = 0$, which shows that $x_\star$ is a solution of $\mathbf{K}x = y$. Moreover,

$$\mathcal{R}(x^\ddagger) \leqslant \mathcal{R}(x_\star) \leqslant \liminf_{n\to\infty} \mathcal{R}(x_{\tau(n)}) \leqslant \liminf_{n\to\infty} \frac{\delta_{\tau(n)}}{\alpha_{\tau(n)}} + \mathcal{R}(x^\ddagger) = \mathcal{R}(x^\ddagger)$$

where for the second last inequality we used (2.4) and for the last equality we used that $\delta_n/\alpha_n \to 0$. Therefore, $x_\star$ is an $\mathcal{R}$-minimizing solution of the equation $\mathbf{K}x = y$. In a similar manner we derive $\mathcal{R}(x_\star) \leqslant \liminf_{n\to\infty} \mathcal{R}(x_{\tau(n)}) \leqslant \limsup_{n\to\infty} \delta_{\tau(n)}/\alpha_{\tau(n)} + \mathcal{R}(x^\ddagger) = \mathcal{R}(x_\star)$ which shows $\mathcal{R}(x_{\tau(n)}) \to \mathcal{R}(x_\star)$.

(d): If $\mathbf{K}x = y$ has a unique $\mathcal{R}$-minimizing solution $x^\ddagger$, then every subsequence of $(x_n)_{n\in\mathbb{N}}$ has itself a subsequence weakly converging to $x^\ddagger$, which implies that $(x_n)_{n\in\mathbb{N}}$ weakly converges to the $\mathcal{R}$-minimizing solution. □

Next we derive strong convergence of the regularized solutions. To this end we recall the absolute Bregman distance, the modulus of total nonlinearity and the total nonlinearity, defined in [31].

Definiton 2.10 Absolute Bregman distance. Let $\mathcal{F}\colon \mathbb{X} \to [0, \infty]$ be Gâteaux differentiable at $x_\star \in \mathbb{X}$. The absolute Bregman distance $\Delta_{\mathcal{F}}(\,\cdot\,, x_\star)\colon \mathbb{X} \to [0, \infty]$ at $x_\star$ with respect to $\mathcal{F}$ is defined by

$$\forall x \in \mathbb{X}: \quad \Delta_{\mathcal{F}}(x, x_\star) := |\mathcal{F}(x) - \mathcal{F}(x_\star) - \mathcal{F}'(x_\star)(x - x_\star)|.$$

Here and below $\mathcal{F}'(x_\star)$ denotes the Gâteaux derivative of $\mathcal{F}$ at $x_\star$.

Definiton 2.11 Modulus of total nonlinearity and total nonlinearity. Let $\mathcal{F}\colon \mathbb{X} \to [0, \infty]$ be Gâteaux differentiable at $x_\star \in \mathbb{X}$. We define the modulus of total nonlinearity of $\mathcal{F}$ at $x_\star$ as $\nu_{\mathcal{F}}(x_\star, \,\cdot\,)\colon [0, \infty) \to [0, \infty)\colon t \mapsto \inf\{\Delta_{\mathcal{F}}(x, x_\star)\colon \|x - x_\star\| = t\}$. We call $\mathcal{F}$ totally nonlinear at $x_\star$ if $\nu_{\mathcal{F}}(x_\star, t) > 0$ for all $t \in (0, \infty)$.

Using these definitions we get the following convergence result in the norm topology.

Theorem 2.12 Strong convergence of aNETT. *Let Conditions 2.1 and 2.4 hold,* $y \in \mathbf{K}(\mathrm{dom}(\mathcal{R}))$ *and let* $\mathcal{R}$ *be totally nonlinear at all* $\mathcal{R}$*-minimizing solutions of* $\mathbf{K}x = y$. *Let* $(y_n)_{n\in\mathbb{N}}$, $(x_n)_{n\in\mathbb{N}}$, $(\alpha_n)_{n\in\mathbb{N}}$ *be as in theorem 2.9. Then there is a subsequence* $(x_{\tau(n)})_{n\in\mathbb{N}}$ *which converges in norm to an* $\mathcal{R}$*-minimizing solution* $x^\ddagger$ *of* $\mathbf{K}x = y$. *If the* $\mathcal{R}$*-minimizing solution is unique, then* $x_n \to x^\ddagger$ *as* $n \to \infty$.

Proof. In [31], proposition 2.9 it is shown that the total nonlinearity of $\mathcal{R}$ implies that for every bounded sequence $(z_n)_{n\in\mathbb{N}}$ with $\Delta_{\mathcal{R}}(z_n, z) \to 0$ it holds that $z_n \to z$. Theorem 2.9 gives us a weakly converging subsequence $(x_{\tau(n)})_{n\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ with weak limit $x^\ddagger$ and $\mathcal{R}(x_{\tau(n)}) \to \mathcal{R}(x^\ddagger)$. By the definition of the absolute Bregman distance it follows that $\Delta_{\mathcal{R}}(x_{\tau(n)}, x^\ddagger) \to 0$ and hence, together with [31], proposition 2.9, that $x_{\tau(n)} \to x^\ddagger$. If the $\mathcal{R}$-minimizing solution of $\mathbf{K}x = y$ is unique, then every subsequence has a subsequence converging to $x^\ddagger$ and hence the claim follows. □

2.4. Convergence rates

We will now prove convergence rates by deriving quantitative estimates for the absolute Bregman distance between $\mathcal{R}$-minimizing solutions for exact data and regularized solutions for noisy data. The convergence rates will be derived under the additional assumption that $\mathcal{D}$ satisfies the quasi triangle-inequality (2.2).

Proposition 2.13 Convergence rates for aNETT. *Let the assumptions of theorem 2.12 be satisfied and suppose that $\mathcal{D}$ satisfies the quasi triangle-inequality (2.2) for some $q \geqslant 1$. Let $x^\ddagger \in \mathbb{X}$ be an $\mathcal{R}$-minimizing solution of $\mathbf{K}x = y$ such that $\mathcal{R}$ is Gâteaux differentiable at $x^\ddagger$ and assume there exist $\epsilon,\ c > 0$ with*

$$\forall x \in \mathbb{X}: \quad \|x - x^\ddagger\| \leqslant \epsilon \Rightarrow \Delta_{\mathcal{R}}(x, x^\ddagger) \leqslant \mathcal{R}(x) - \mathcal{R}(x^\ddagger) + c\sqrt{\mathcal{D}(\mathbf{K}x, \mathbf{K}x^\ddagger)}. \tag{2.5}$$

For any $\delta > 0$, let $y^\delta \in \mathbb{Y}$ be noisy data satisfying $\mathcal{D}(y^\delta, \mathbf{K}x^\ddagger)^{1/2} \leqslant \delta$, $\mathcal{D}(\mathbf{K}x^\ddagger, y^\delta)^{1/2} \leqslant \delta$, and write $x_\alpha^\delta \in \arg\min \mathcal{A}_{\alpha,y^\delta}$. Then the following hold:

(a) For sufficiently small α, it holds $\Delta_{\mathcal{R}}(x_\alpha^\delta, x^\ddagger) \leqslant \delta^2/\alpha - c\delta\sqrt{q} + c^2\alpha q/4$.

(b) If $\alpha \asymp \delta$, then $\Delta_{\mathcal{R}}(x_\alpha^\delta, x^\ddagger) = \mathcal{O}(\delta)$ as $\delta \to 0$.

Proof. By definition of x_α^δ we have $\mathcal{D}(\mathbf{K}x_\alpha^\delta, y^\delta) + \alpha\mathcal{R}(x_\alpha^\delta) - \alpha\mathcal{R}(x^\ddagger) \leqslant \mathcal{D}(\mathbf{K}x^\ddagger, y^\delta) \leqslant \delta^2$. By theorem 2.12 for sufficiently small α we can assume that $\|x_\alpha^\delta - x^\ddagger\| \leqslant \epsilon$ and hence

$$\begin{aligned}
\alpha\Delta_{\mathcal{R}}(x_\alpha^\delta, x^\ddagger) &\leqslant \alpha\mathcal{R}(x_\alpha^\delta) - \alpha\mathcal{R}(x^\ddagger) + c\alpha\sqrt{\mathcal{D}(\mathbf{K}x_\alpha^\delta, \mathbf{K}x^\ddagger)} \\
&= \mathcal{R}(x_\alpha^\delta) - \mathcal{D}(\mathbf{K}x_\alpha^\delta, y^\delta) - (\mathcal{R}(x_\alpha^\delta) - \mathcal{D}(\mathbf{K}x_\alpha^\delta, y^\delta)) + c\alpha\sqrt{\mathcal{D}(\mathbf{K}x_\alpha^\delta, \mathbf{K}x^\ddagger)} \\
&\leqslant \delta^2 - \mathcal{D}(\mathbf{K}x_\alpha^\delta, y^\delta) + c\alpha\delta\sqrt{q} + c\alpha\sqrt{q\mathcal{D}(\mathbf{K}x_\alpha^\delta, y^\delta)}.
\end{aligned}$$

Together with the inequality of arithmetic and geometric means $(a+b)/2 \geqslant \sqrt{ab}$ for $a = \mathcal{D}(\mathbf{K}x_\alpha^\delta, y^\delta)$ and $b = c^2\alpha^2 q/4$ this implies $\alpha\Delta_{\mathcal{R}}(x_\alpha^\delta, x^\ddagger) \leqslant \delta^2 - c\alpha\delta\sqrt{q} + c^2\alpha^2 q/4$ which shows (a). Item (b) is an immediate consequence of (a). □

The following results is our main convergence rates result. It is similar to proposition [31], theorem 3.1, but uses different assumptions.

Theorem 2.14 Convergence rates for finite rank operators. *Let the assumptions of theorem 2.12 be satisfied, take $\mathcal{D}(y_1, y_2) = \|y_1 - y_2\|^2$, assume $\mathbf{K}$ has finite dimensional range and that $\mathcal{R}$ is Lipschitz continuous and Gâteaux differentiable. For any $\delta > 0$, let $y^\delta \in \mathbb{Y}$ be noisy data satisfying $\|y^\delta - \mathbf{K}x^\ddagger\| \leqslant \delta$ and write $x_\alpha^\delta \in \arg\min \mathcal{A}_{\alpha,y^\delta}$. Then for the parameter choice $\alpha \asymp \delta$ we have the convergence rates result $\Delta_{\mathcal{R}}(x_\alpha^\delta, x^\ddagger) = \mathcal{O}(\delta)$ as $\delta \to 0$.*

Proof. According to proposition 2.13, it is sufficient to show that (2.5) holds with $\|\mathbf{K}x - \mathbf{K}x^\ddagger\|$ in place of $\mathcal{D}(\mathbf{K}x, \mathbf{K}x^\ddagger)^{1/2}$. For that purpose, let $\mathbf{P}$ denote the orthogonal projection onto the null-space $\ker(\mathbf{K})$ and let L be a Lipschitz constant of $\mathcal{R}$. Since $\mathbf{K}$ restricted to $\ker(\mathbf{K})^\perp$ is injective with finite dimensional range, we can choose a constant $a > 0$ such that $\forall z \in \ker(\mathbf{K})^\perp$: $\|\mathbf{K}z\| \geqslant a\|z\|$.

We first show the estimates

$$\forall x \in \mathbb{X}: \quad \mathcal{R}(x^\ddagger) - \mathcal{R}(x) \leqslant (L/a)\,\|\mathbf{K}x - \mathbf{K}x^\ddagger\| \tag{2.6}$$

$$\forall x \in \mathbb{X}: \quad |\langle \mathcal{R}'(x^\ddagger), x - x^\ddagger\rangle| \leqslant (\|\mathcal{R}'(x^\ddagger)\|/a)\,\|\mathbf{K}x - \mathbf{K}x^\ddagger\|. \tag{2.7}$$

To that end, let $x \in \mathbb{X}$ and write $x_0 := (x^\ddagger - \mathbf{P}x^\ddagger) + \mathbf{P}x$. Then $\mathbf{K}x_0 = \mathbf{K}x^\ddagger$. Since $x^\ddagger$ is an $\mathcal{R}$-minimizing solution, we have $\mathcal{R}(x^\ddagger) - \mathcal{R}(x) \leqslant \mathcal{R}(x_0) - \mathcal{R}(x) \leqslant L\|x_0 - x\|$. Since $x_0 - x \in \ker(\mathbf{K})^\perp$, we have $\|\mathbf{K}x^\ddagger - \mathbf{K}x\| = \|\mathbf{K}(x_0 - x)\| \geqslant a\|x_0 - x\|$. The last two estimates prove (2.6). Because $x^\ddagger$ is an $\mathcal{R}$-minimizing solution, we have $\langle \mathcal{R}'(x^\ddagger), x^\ddagger - x\rangle = 0$ whenever $x^\ddagger - x \in \ker(\mathbf{K})$. On the other hand, using that $\mathcal{R}$ is Gâteaux differentiable and that $\mathbf{K}$ has finite rank, shows $|\langle \mathcal{R}'(x^\ddagger), x^\ddagger - x\rangle| \leqslant \|\mathcal{R}'(x^\ddagger)\|\, a^{-1}\|\mathbf{K}(x^\ddagger - x)\|$ for $x^\ddagger - x \in \ker(\mathbf{K})^\perp$. This proves (2.7).

Inequality (2.6) implies $|\mathcal{R}(x) - \mathcal{R}(x^\ddagger)| \leqslant \mathcal{R}(x) - \mathcal{R}(x^\ddagger) + 2\,(L/a)\,\|\mathbf{K}x - \mathbf{K}x^\ddagger\|$. Together with (2.7) this yields

$$\begin{aligned}
\Delta_{\mathcal{R}}(x, x^\ddagger) &= |\mathcal{R}(x) - \mathcal{R}(x^\ddagger) - \langle \mathcal{R}'(x^\ddagger), x - x^\ddagger\rangle| \\
&\leqslant |\mathcal{R}(x) - \mathcal{R}(x^\ddagger)| + |\langle \mathcal{R}'(x^\ddagger), x^\ddagger - x\rangle| \\
&\leqslant \mathcal{R}(x) - \mathcal{R}(x^\ddagger) + (2L + \|\mathcal{R}'(x^\ddagger)\|)\, a^{-1}\,\|\mathbf{K}x^\ddagger - \mathbf{K}x\|,
\end{aligned}$$

which proves (2.5) with $c = (2L + \|\mathcal{R}'(x^\ddagger)\|)/a$. □

Note that the theoretical results stated remain valid, if we replace $\mathcal{R}$ by a general coercive and weakly lower semi-continuous regularizer $\mathcal{R}\colon \mathbb{X} \to [0, \infty]$.

3. Practical realization

In this section we investigate practical aspects of aNETT. We present a possible network architecture together with a possible training strategy in the discrete setting[3]. Further we discuss minimization of aNETT using the ADMM algorithm. For the sake of clarity we restrict our discussion to the finite dimensional case where $\mathbb{X} = \mathbb{R}^{N\times N}$ and $\ell^2(\Lambda) = \mathbb{R}^{\Lambda}$ for a finite index set Λ.

3.1. Proposed modular aNETT training

To find a suitable network $\mathbf{D}\circ\mathbf{E}$ defining the aNETT regularizer $\mathcal{R}(x) = \mathcal{Q}(\mathbf{E}(x)) + (c/2)\,\|x - \mathbf{D}\circ\mathbf{E}(x)\|^2$, we propose a modular data driven approach that comes in two separate steps. In a first step, we train a $\mathcal{Q}$-regularized denoising autoencoder $\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E}$ independent of the forward problem $\mathbf{K}$, whose purpose is to well represent elements of a training data set by low complexity encoder coefficients. In a second step, we train a task-specific network that increases the ability of the aNETT regularizer to distinguish between clean images and images containing problem specific artifacts.

Let $x_1, \ldots, x_m \in \mathcal{M}$ denote the given set of artifact-free training phantoms.

- $\mathcal{Q}$–REGULARIZED AUTOENCODER:

First, an autoencoder $\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E}$ is trained such that x_i is close to $\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E}(x_i)$ and that $\mathcal{Q}(\mathbf{E}(x_i))$ is small for the given training signals. For that purpose, let $(\mathbf{D}_\theta\circ\mathbf{E}_\theta)_{\theta\in\Theta}$ be a family of autoencoder networks, where $\mathbf{E}_\theta\colon \mathbb{R}^{N\times N} \to \mathbb{R}^{\Lambda}$ are encoder and $\mathbf{D}_\theta\colon \mathbb{R}^{\Lambda} \to \mathbb{R}^{N\times N}$ decoder networks, respectively.

To achieve that unperturbed images are sparsely represented by $\mathbf{E}$, whereas disrupted images are not, we apply the following training strategy. We randomly generate images $x_i + a_i\epsilon_i$ where ϵ_i is additive Gaussian white noise with a standard deviation proportional to the mean value of x_i, and $a_i \in \{0, 1\}$ is a binary random variable that takes each value with probability 0.5. For the numerical results below we use a standard deviation of 0.05 times the mean value of x_i. To select the particular autoencoder based on the training data, we consider the following training strategy

$$\theta^* \in \arg\min_\theta \frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}} \|(\mathbf{D}_\theta\circ\mathbf{E}_\theta)(\mathrm{x}_i + \mathrm{a}_i\epsilon_i) - \mathrm{x}_i\|_2^2 + \nu(1 - \mathrm{a}_i)\,\mathcal{Q}(\mathbf{E}_\theta(\mathrm{x}_i)) + \beta\,\|\theta\|_2^2, \tag{3.1}$$

and set $[\mathbf{D}^{\mathcal{Q}}, \mathbf{E}] := [\mathbf{D}_{\theta^*}, \mathbf{E}_{\theta^*}]$. Here $\nu, \beta > 0$ are regularization parameters.

Including perturbed signals $x_i + \epsilon_i$ in (3.1) increases robustness of the $\mathcal{Q}$-regularized autoencoder. To enforce regularity for the encoder coefficients only on the noise-free images, the penalty $\mathcal{Q}(\mathbf{E}(\,\cdot\,))$ is only used for the noise-free inputs, reflected by the pre-factor $1 - a_i$. Using auto-encoders, regularity for a signal class could also be achieved by means of dimensionality reduction techniques, where $\mathbb{R}^{\Lambda}$ is used as a bottleneck in the network architecture. However, in order to get a regularizer that is able to distinguish between perturbed and unperturbed signals we use $\mathbb{R}^{\Lambda}$ to be of sufficiently high dimensionality.

- TASK-SPECIFIC NETWORK:

Numerical simulations showed that the $\mathcal{Q}$-regularized autoencoder alone was not able to sufficiently well distinguish between artifact-free training phantoms and images containing problem specific artifacts. In order to address this issue, we compose the operator independent network with another network $\mathbf{U}$, that is trained to distinguish between images with and without problem specific artifacts.

For that purpose, we consider randomly generated images $z_1, \ldots, z_m$ where either $z_i = (\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E})(x_i)$ or $z_i = (\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E})(\mathbf{K}^{\ddagger}(\mathbf{K}x_i + \eta_i))$ with equal probability. Here $\mathbf{K}^{\ddagger}$ is an approximate right inverse and η_i are error terms We choose a network architecture $(\mathbf{U}_\theta)_{\theta\in\Theta}$ and select $\mathbf{U} = \mathbf{U}_{\theta^*}$, where

$$\theta^* \in \arg\min \frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}} \|\mathbf{U}_\theta(\mathrm{z}_i) - \mathrm{x}_i\|_2^2 + \gamma\,\|\theta\|_2^2, \tag{3.2}$$

for some regularization parameter $\gamma > 0$. In particular, the image residuals $(\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E})(\mathbf{K}^{\ddagger}(\mathbf{K}x_i + \eta_i)) - (\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E})(x_i)$ now depend on the specific inverse problem and we can consider them to consist of operator and training signal specific artifacts.

[3] https://git.uibk.ac.at/c7021101/augmented-nett.git

Table 1. Parameter specifications for proposed aNETT functional and it numerical minimization.

	α	c	γ	N_{iter}	N_φ	noise model	$\mathcal{D}$
Sparse view	10^{-4}	10^2	$5 \cdot 10^{-1}$	50	40	Gaussian, $\sigma = 0.02$	$\mathcal{D}_{\|\cdot\|^2}$
Low dose	$5 \cdot 10^{-3}$	10^2	10^{-3}	20	1138	Poisson, $p = 10^4$	$\mathcal{D}_{\text{KL}}$
Universality	10^{-4}	10^2	$5 \cdot 10^{-1}$	50	160	Gaussian, $\sigma = 0.02$	$\mathcal{D}_{\|\cdot\|^2}$

The above training procedure ensures that the network $\mathbf{U}$ adapts to the inverse problem at hand as well as to the $\mathcal{Q}$-regularized autoencoder. Training the network $\mathbf{U}$ independently of $\mathbf{D}^{\mathcal{Q}}\circ\mathbf{E}$, or directly training the autoencoder to distinguish between images with and without problem specific artifacts, we empirically found to perform considerably worse.

The final autoencoder is then given as $\mathbf{N} = \mathbf{D}\circ\mathbf{E}$ with modular decoder $\mathbf{D} := \mathbf{U}\circ\mathbf{D}^{\mathcal{Q}}$. For the numerical results we take $(\mathbf{U}_\theta)_{\theta\in\Theta}$ as the tight frame U-Net of [38]. Moreover, we choose $(\mathbf{D}_\theta\circ\mathbf{E}_\theta)_{\theta\in\Theta}$ as modified tight frame U-Net proposed in [37] for deep synthesis regularization. In particular, opposed to the original tight frame U-net, the modified tight frame U-Net does not involve skip connections.

3.2. Possible aNETT minimization

For minimizing the aNETT functional (1.3) we use the alternating direction method of multiplies (ADMM) with scaled dual variable [41–43]. For that purpose, the aNETT minimization problem is rewritten as the following constraint minimization problem

$$\begin{cases} \arg\min_{x,\xi} & \mathcal{D}(\mathbf{K}x,\, y^\delta) + \alpha\mathcal{Q}(\xi) + \alpha c/2 \;\; \|x - \mathbf{N}(x)\|^2 \\ \text{s.t.} & \mathbf{E}(x) = \xi. \end{cases}$$

The resulting ADMM update scheme with scaling parameter $\rho > 0$ initialized by $\xi_0 = \mathbf{E}(\mathbf{K}^\ddagger y^\delta)$ and $\eta_0 = 0$ then reads as follows:

(S1) $x_{k+1} = \arg\min_x \mathcal{D}(\mathbf{K}\mathrm{x},\, \mathrm{y}^\delta) + (\alpha\mathrm{c}/2)\, \|\mathrm{x} - \mathbf{N}(\mathrm{x})\|_2^2 + (\rho/2)\, \|\mathbf{E}(\mathrm{x}) - \xi_k + \eta_k\|_2^2$.
(S2) $\xi_{k+1} = \arg\min_\xi \alpha\mathcal{Q}(\xi) + (\rho/2)\, \|\mathbf{E}(\mathrm{x}_{k+1}) - \xi + \eta_k\|_2^2$.
(S3) $\eta_{k+1} = \eta_k + (\mathbf{E}(x_{k+1}) - \xi_{k\,+\,1})$.

One interesting feature of the above approach is that the signal update (S1) is independent of the possibly non-smooth penalty $\mathcal{Q}$. Moreover, the encoder update (S2) uses the proximal mapping of $\mathcal{Q}$ which in important special cases can be evaluated explicitly and therefore fast and exact. Moreover, it guarantees regular encoder coefficients during each iteration. For example, if we choose the penalty as the ℓ^1-norm, then (S2) is a soft-thresholding step which results in sparse encoder coefficients. Step (S1) in typical cases has to be computed iteratively via an inner iteration. To find an approximate solution for (S1) for the results presented below we use gradient descent with at most 10 iterations. We stop the gradient descent updates early if the difference of the functional evaluated at two consecutive iterations is below our predefined tolerance of 10^{-5}.

The concrete implementation of the aNETT minimization requires specification of the similarity measure, the total number of outer iterations N_{iter}, the step-size γ for the iteration in (S1) and the parameters defining the aNETT functional. These specifications are selected dependent of the inverse problem at hand. Table 1 lists the particular choices for the reconstruction scenarios considered in the following section.

In order to choose the parameters for the numerical simulations we have tested different values and manually chose the parameters which maximized performance among the considered parameters. Another way of choosing these parameters could be to try and learn these parameters from the data using some kind of machine learning approach or choose a bilevel approach similar to [44].

In the simulations we have observed that choosing c larger will tend to oversmooth the resulting reconstructions. Taking a smaller value for c we observed that the manifold term $\|x - \mathbf{N}(x)\|^2$ tends to be undervalued resulting in worse performance. In a similar fashion we found that choosing α larger will have a smoothing effect on the resulting reconstructions while lowering α will make the reconstructions less smooth.

The ADMM scheme for aNETT minimization shares similarities with existing iterative neural network based reconstruction methods. In particular, ADMM inspired plug-and-play priors [15–17] may be most closely related. However, opposed the plug and play approach we can deduce convergence from existing results for ADMM for non-convex problems [45]. While convergence of (S1)–(S3) and relations with plug and play priors are interesting and relevant, they are beyond the scope of this work. This also applies to the comparison with other iterative minimization schemes for minimizing aNETT.

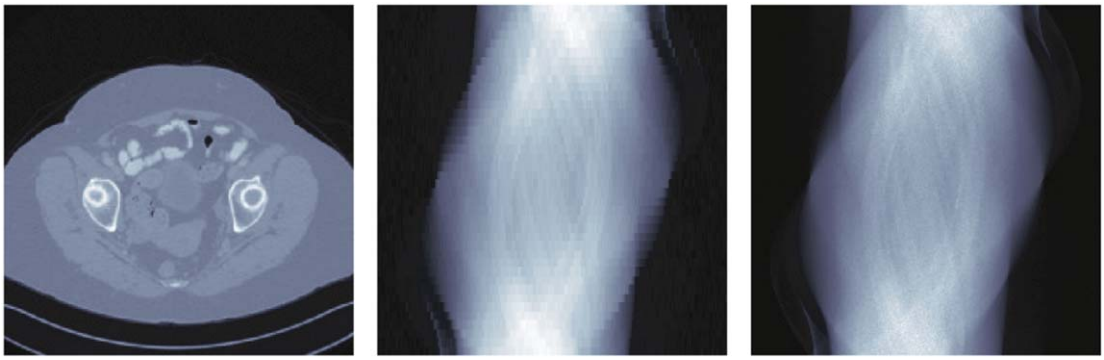

Figure 1. Left: Example image randomly drawn from the dataset. Middle: Corresponding sparse view sinogram (40 directions). Right: low dose sinogram.

4. Application to sparse view and low dose CT

In this section we apply aNETT regularization to sparse view and low-dose computed tomography (CT). For the experiments we always choose $\mathcal{Q}$ to be the ℓ^1-norm. The parameter specifications for the proposed aNETT functional and its numerical minimization are given in table 1. For quantitative evaluation, we use the peak-signal-to-noise-ratio (PSNR) defined by

$$\mathrm{PSNR}(x,\, x_{\mathrm{rec}}) \coloneqq 20 \log_{10}\left(\frac{\max x}{\|x - x_{\mathrm{rec}}\|_2}\right).$$

Here $x \in \mathbb{R}^{N\times N}$ is the ground truth image and $x_{\mathrm{rec}} \in \mathbb{R}^{N\times N}$ its numerical reconstruction. Higher value of PSNR indicates better reconstruction.

4.1. Discretization and dataset

For sparse view CT as well as for low dose CT we work with a discretization of the Radon transform $(\mathbf{R}f)(\varphi,\, s) \coloneqq \int_{\mathbb{R}} f(s\cos(\varphi) - t\sin(\varphi),\, s\sin(\varphi) + t\cos(\varphi))\mathrm{d}t$. The values $(\mathbf{R}f)(\varphi,\, s)$ are integrals of the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ over lines orthogonal to $(\cos(\varphi),\, \sin(\varphi))^\intercal$ for angle $\varphi \in [0,\, \pi)$ and signed distance $s \in \mathbb{R}$. We discretize the Radon transform using the ODL library [46] where we assume that the function has compact support in $[-1,\, 1]^2$ and sampled on an equidistant grid. We use N_φ equidistant samples of $\varphi \in [0,\, \pi)$ and N_s equidistant samples of $s \in [-1.5,\, 1.5]$. In both cases, we end up with an inverse problem of the form (1.1), where $\mathbf{K}\colon \mathbb{R}^{N\times N} \to \mathbb{R}^{N_\varphi \times N_s}$ is the discretized linear forward operator. Elements $x \in \mathbb{R}^{N\times N}$ will be referred to as CT images and the elements $y \in \mathbb{R}^{N_\varphi \times N_s}$ as sinograms.

For all results presented below we work with image size 512×512 and use $N_s = 768$. The number of angular samples N_φ is taken 40 for sparse view CT and $N_\varphi = 1138$ for the low dose example. In both cases we use the CT images from the Low Dose CT Grand Challenge dataset [47] provided by the Mayo Clinic. The dataset consists of 512×512 grayscale images of 10 different patients, where for each patient there are multiple CT scanning series available. We use the split $7/2/1$ for training, validation and testing which corresponds to $4267/1143/526$ CT images in the respective sets. We use the validation set to select networks which achieve the minimal loss on the validation set. The test set is used to evaluate the final performance. Note that by splitting of the dataset according to patient we avoid validation and testing on images patients that have already be seen during training time. An example image and the corresponding simulated sparse view and low-dose sinogram are shown in figure 1.

4.2. Numerical results

We compare results of aNETT to the learned primal-dual algorithm (LPD) [48], the tight frame U-Net [38] applied as post-processing network (CNN) and the filtered back-projection (FBP). Minimization of the loss-function for all methods was done using Adam [49] for 100 epochs, cosine decay learning rate $\eta_t = (\eta_0/2)\cdot(1 + \cos(\pi t/100))$ with $\eta_0 = 10^{-3}$ in the t-th epoch, and a batch-size of 4. For LPD we take the hyper-parameters $N_{\mathrm{primal}} = N_{\mathrm{dual}} = 5$ and $N = 7$ network iterations and train according to [48]. Here, we choose to only use $N = 7$ network iterations because we observed instabilities during the training phase when this parameter was chosen larger and we have not performed any parametr tuning. For training of the tight frame U-Net we do not follow the patch approach of [38] but instead use full images obtained with FBP as CNN inputs. Training of all the networks was done on a GTX 1080 Ti with an Intel Xeon Bronze 3104 CPU.

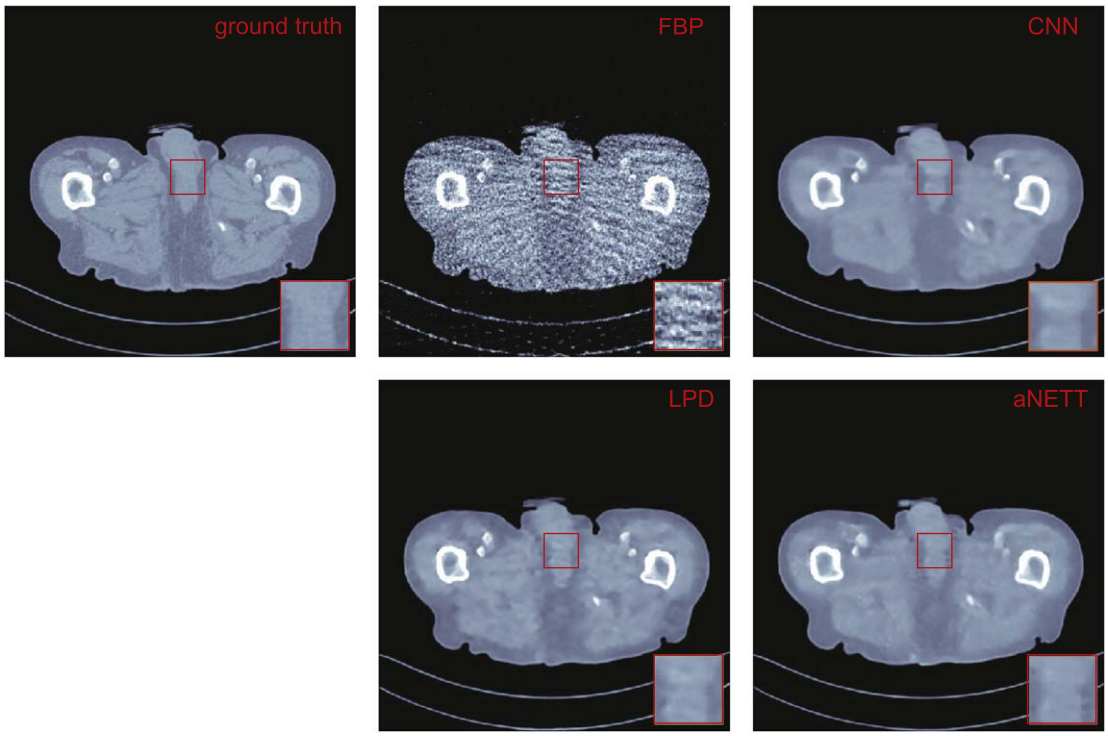

Figure 2. Reconstructions for sparse view CT data from $N_\varphi = 40$ angular directions. The intensity range of all images is $[-500, 500]$ HU.

Table 2. Overview of metric results evaluated on the test-set. The values shown are the average of the PSNR $\pm$ the standard deviation calculated over the test dataset. The values in bold show the best results. The `na` entry means that LPD was not applied to this problem setting, as in the used framework there is no canonical way to use LPD with modified sampling pattern.

PSNR	FBP	LPD	Post	aNETT
Sparse view	23.8 $\pm$ 1.3	**37.9 $\pm$ 1.2**	37.1 $\pm$ 0.9	37.1 $\pm$ 1.0
Low dose	36.9 $\pm$ 1.6	43.6 $\pm$ 1.3	**44.1 $\pm$ 1.5**	43.9 $\pm$ 1.3
Universality	32.4 $\pm$ 1.6	na	37.7 $\pm$ 0.8	**38.3 $\pm$ 1.0**

- SPARSE VIEW CT: to simulate sparse view data we evaluate the Radon transform for $N_\varphi = 40$ directions. We generate noisy data $y^\delta = \mathbf{K}x + \eta^\delta$ by adding Gaussian white noise with standard deviation taken as 0.02 times the mean value of $\mathbf{K}x$. We use the ℓ^2-norm distance as the similarity measure. Quantitative results evaluated on the test set are shown in table 2. All learning-based methods yield comparable performance in terms of PSNR and clearly outperform FBP. The reconstructions shown in figure 2 indicate that aNETT reconstructions are less smooth than CNN reconstructions and less blocky than LPD reconstructions.
- LOW DOSE CT: for the low dose problem, we use a fully sampled sinogram with $N_\varphi = 1138$ and add Poisson noise corresponding to 10^4 incident photons per pixel bin. The Kullback-Leibler divergence $\mathcal{D}_{\mathrm{KL}}$ is a more appropriate discrepancy term than the squared ℓ^2-norm distance in case of Poisson noise and the reported values and reconstructions use the Kullback-Leibler divergence as the similarity measure. Quantitative results are shown in table 2. Again, all learning-based methods give similar results and significantly outperform FBP. Visual comparison of the reconstructions in figure 3 shows that CNN yields cartoon like images and the LPD reconstruction again looks blocky. The aNETT reconstruction shows more texture than the CNN reconstruction and at the same time is less blocky than the LPD reconstruction.
- UNIVERSALITY: in practical applications, we may not have a fixed sampling pattern. If we have many different sampling patterns, then training a network for each sampling pattern is infeasible and hence reconstruction methods should be applicable to different sampling scenarios. Additionally, it is desirable that an increased number of samples indeed increases performance. In order to test this issue, we consider the sparse view CT problem but with an increased number of angular samples without retraining the networks. Due to the rigidity

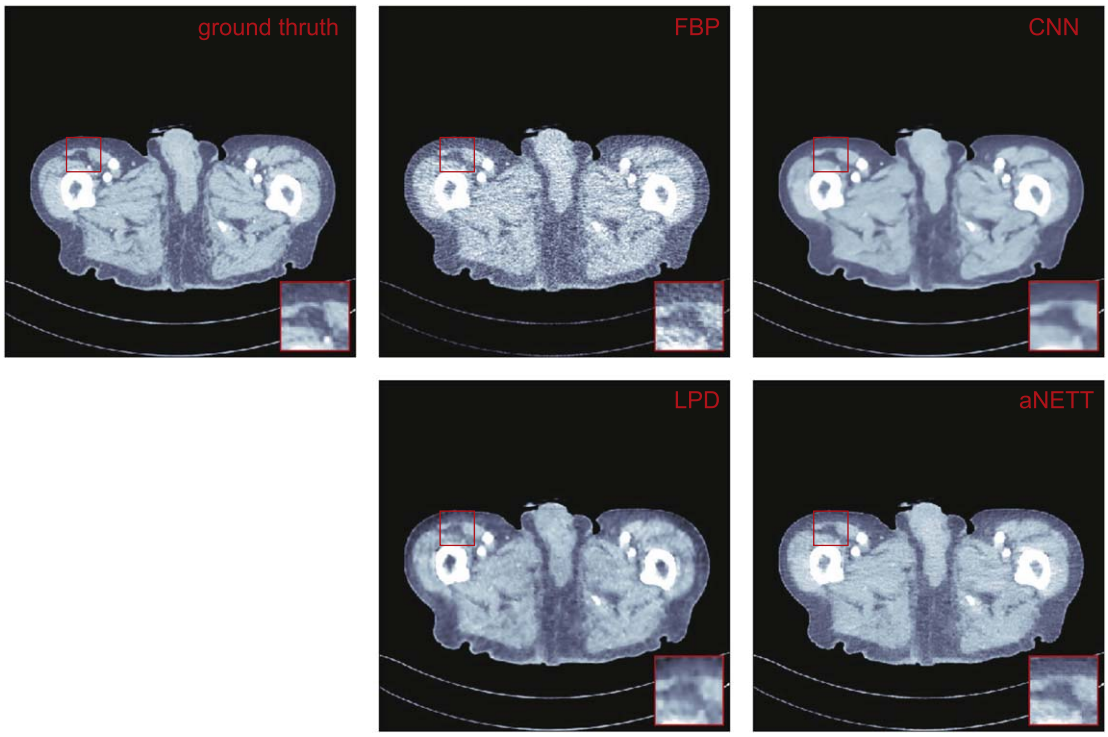

Figure 3. Reconstructions results from low dose CT data. The intensity range of all images is $[-200,\ 200]$ HU.

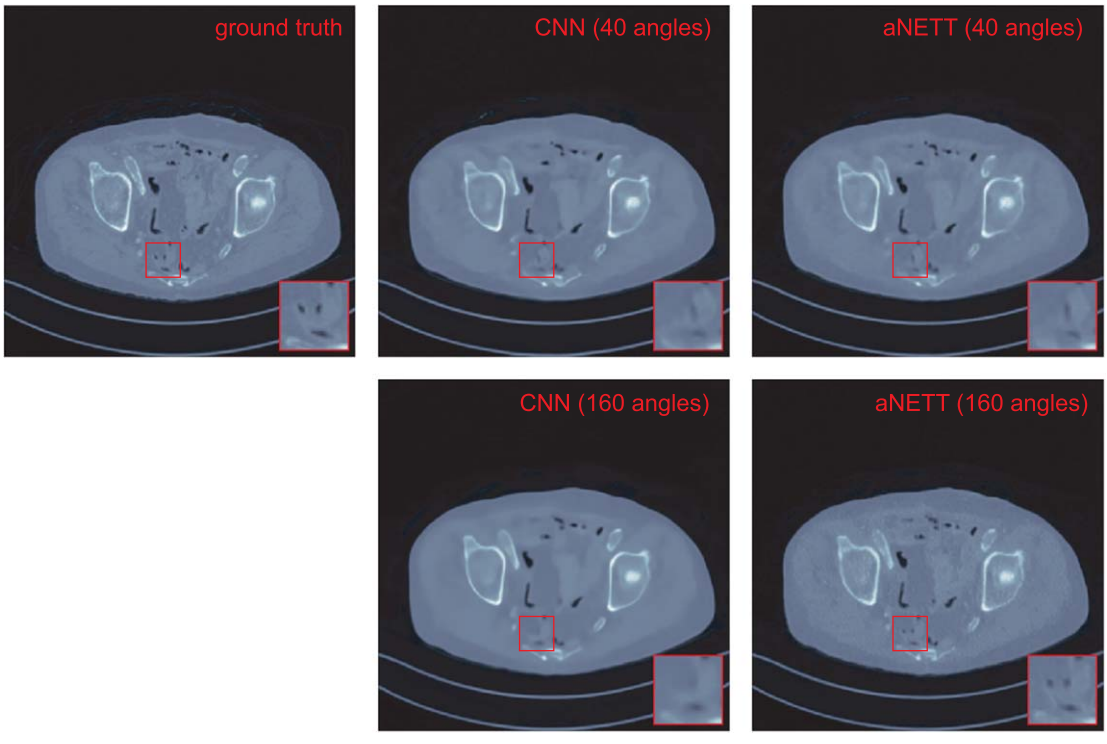

Figure 4. Universality of aNETT due to change of angular sampling pattern. Top row: Ground truth and reconstructions from 40 angular directions. Bottom: Reconstructions from 160 angular directions. All reconstruction use the networks trained with 40 angular directions. While aNETT shows increased resolution for ncreased angular sampling, CNN does not.

of the used framework LPD cannot easily be adapted to this problem and we therefore decided to only compare aNETT with the post-processing CNN. For the results presented here, no network was retrained. Quantitative evaluation for this scenario is given in table 2. We see that aNETT slightly outperforms the CNN in terms of PSNR. The advantage of aNETT over CNN, however, is best observed in figure 4. One observes

that CNN yields a similar reconstructions for both angular sampling patterns. On the other hand, aNETT is able to synergistically combine the increased sampling rate of the sinogram with the network trained on coarsely sampled data. Despite using the network trained with only 40 angular samples, aNETT reconstructs small details which are not present in the reconstruction from 40 angular samples.

4.3. Discussion

The results show that the proposed aNETT regularization is competitive with prominent deep-learning methods such as LPD and post-processing CNNs. We found that the aNETT does not suffer as much from over-smoothing which is often observed in other deep-learning reconstruction methods. This can for example be seen in figure 3 where the CNN yields an over-smoothed reconstruction and the aNETT reconstruction shows more texture. Besides this, aNETT reconstructions are less blocky than LPD reconstructions. Moreover, aNETT is able to leverage higher sampling rates without retraining the networks to reconstruct small details while other deep-learning methods fail to do so. We conjecture that this advantage arises due to the fact that aNETT can make use of the higher sampling rate using the data-consistency term in (1.3), while the CNN is agnostic to this change in the sampling rate. In some scenarios, it may not be possible to retrain networks. Especially for learned iterative schemes network training is a time-consuming task. Training aNETT on the other hand is straightforward and, as demonstrated, yields a method which is robust to changes of the forward problem during testing time.

While a more extensive study with respect to the influence of noise could be done to further analyse the advantages and disadvantages of each method, this is not our main focus here and is thus postponed to a future study.

Finally, we note that aNETT relies on minimizing (1.3) iteratively. With the use of the ADMM minimization scheme presented in this article, aNETT is slower than the methods used for comparison in this article. Designing faster optimization schemes for (1.3) is beyond the scope of this work, but is an important and interesting aspect.

5. Conclusion

We have proposed the aNETT (augmented NETwork Tikhonov) for which we derived coercivity of the regularizer under quite mild assumptions on the networks involved. Using this coercivity we presented a convergence analysis of aNETT with a general similarity measure $\mathcal{D}$. We proposed a modular training strategy in which we first train an $\mathcal{Q}$-regularized autoencoder independent of the problem at hand and then a network which is adapted to the problem and first autoencoder. Experimentally we found this training strategy to be superior to directly training the autoencoder on the full task. Lastly, we conducted numerical simulations demonstrating the feasibility of aNETT.

The experiments show that aNETT is able to keep up with the classical post-processing CNNs and the learned primal-dual approach for sparse view and low dose CT. Typical deep learning methods work well for a fixed sampling pattern on which they have been trained on. However, reconstruction methods are expected to perform better if we use an increased sampling rate. We have experimentally shown that aNETT is able to leverage higher sampling rates to reconstruct small details in the images which are not visible in the other reconstructions. This universality can be advantageous in applications where one is not fixed to one sampling pattern or is not able to train a network for every sampling pattern.

Acknowledgments

D O and M H acknowledge support of the Austrian Science Fund (FWF), project P 30 747-N32. The research of L N has been supported by the National Science Foundation (NSF) Grants DMS 1 212 125 and DMS 1 616 904.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Daniel Obmann https://orcid.org/0000-0002-7130-5464
Linh Nguyen https://orcid.org/0000-0003-0776-9480

References

[1] Engl H W, Hanke M and Neubauer A 1996 Regularization of inverse problems *Mathematics and its Applications* vol 375 (Dordrecht: Kluwer Academic Publishers Group)
[2] Scherzer O, Grasmair M, Grossauer H, Haltmeier M and Lenzen F 2009 Variational methods in imaging *Applied Mathematical Sciences* vol 167 (New York: Springer)
[3] Arridge S, Maass P, Öktem O and Schönlieb C-B 2019 Solving inverse problems using data-driven models *Acta Numer.* **28** 1–174
[4] Haltmeier M and Nguyen L V 2020 Regularization of inverse problems by neural networks arXiv:2006.03972
[5] Maier A, Syben C, Lasser T and Riess C 2019 A gentle introduction to deep learning in medical image processing *Zeitschrift für Medizinische Physik* **29** 86–101
[6] McCann M T, Jin K H and Unser M 2017 Convolutional neural networks for inverse problems in imaging: A review *IEEE Signal Process. Mag.* **34** 85–95
[7] Wang G 2016 A perspective on deep imaging *IEEE Access* **4** 8914–24
[8] Antholzer S, Haltmeier M and Schwab J 2018 Deep learning for photoacoustic tomography from sparse data *Inverse Probl. Sci. and Eng.* **27** 987–1005
[9] Jin K H, McCann M T, Froustey E and Unser M 2017 Deep convolutional neural network for inverse problems in imaging *IEEE Trans. Image Process* **26** 4509–22
[10] Lee D, Yoo J and Ye J C 2017 Deep residual learning for compressed sensing MRI *IEEE XIV International Symposium on Biomedical Imaging (Melbourne, VIC, Australia, 18–21 April 2017)* (Picastaway, NJ: IEEE) pp 15–8
[11] Maier A K, Syben C, Stimpel B, Würfl T, Hoffmann M, Schebesch F, Fu W, Mill L, Kling L and Christiansen S 2019 Learning with known operators reduces maximum error bounds *Nat. Mach. Intell.* **1** 373–80
[12] Rivenson Y, Zhang Y, Günaydın H, Teng D and Ozcan A 2018 Phase recovery and holographic image reconstruction using deep learning in neural networks *Light Sci. Appl.* **7** 17141–17141
[13] Schwab J, Antholzer S and Haltmeier M 2019 Deep null space learning for inverse problems: convergence analysis and rates *Inverse Probl.* **35** 025008
[14] Schwab J, Antholzer S and Haltmeier M 2020 Big in Japan: Regularizing networks for solving inverse problems *J. Math. Imaging Vis.* **62** 445–55
[15] Chan S H, Wang X and Elgendy O A 2016 Plug-and-play admm for image restoration: Fixed-point convergence and applications *IEEE Trans. Comput. Imag.* **3** 84–98
[16] Romano Y, Elad M and Milanfar P 2017 The little engine that could: Regularization by denoising (red) *SIAM J. Imaging Sci.* **10** 1804–44
[17] Venkatakrishnan S V, Bouman C A and Wohlberg B 2013 Plug-and-play priors for model based reconstruction *2013 IEEE Global Conf. on Signal and Information Processing (Austin, TX, 3–5 December 2013)* (Piscataway, NJ: IEEE) pp 945–8
[18] Dittmer S, Kluth T, Maass P and Baguer D O 2020 Regularization by architecture: A deep prior approach for inverse problems *J. Math. Imaging Vis.* **62** 456–70
[19] Ulyanov D, Vedaldi A and Lempitsky V 2018 Deep image prior *Proceedings of the IEEE conference on Computer vision and Pattern Recognition* 9446–54
[20] Hammernik K, Klatzer T, Kobler E, Recht M P, Sodickson D K, Pock T and Knoll F 2018 Learning a variational network for reconstruction of accelerated mri data *Magn. Reson. Med.* **79** 3055–71
[21] Kobler E, Klatzer T, Hammernik K and Pock T 2017 Variational networks: connecting variational methods and deep learning *German Conference on Pattern Recognition* 281–93
[22] Kofler A, Haltmeier M, Kolbitsch C, Kachelrieß M and Dewey M 2018 A U-Nets cascade for sparse view computed tomography. In International Workshop on *Machine Learning for Medical Image Reconstruction (MLIMR)* (Berlin: Springer) 91–9
[23] Schlemper J, Caballero J, Hajnal J V, Price A and Rueckert D 2017 A deep cascade of convolutional neural networks for mr image reconstruction *Proc. Inf. Process. Med. Imaging* (Berlin: Springer) 647–58
[24] Adler J and Öktem O 2017 Solving ill-posed inverse problems using iterative deep neural networks *Inverse Probl.* **33** 124007
[25] Aggarwal H K, Mani M P and Jacob M 2018 MoDL: model-based deep learning architecture for inverse problems *IEEE Trans. Med. Imag.* **38** 394–405
[26] Chang J R, Li C-L, Poczos B and Kumar B V 2017 One network to solve them all-solving linear inverse problems using deep projection models *IEEE International Conference on Computer Vision (ICCV)* 5889–98
[27] Hauptmann A, Adler J, Arridge S R and Oktem O 2021 Multi-scale learned iterative reconstruction *IEEE Trans. Comput. Imag.* **6** 843–56
[28] Kofler A, Haltmeier M, Schaeffter T and Kolbitsch C 2021 An end-to-end-trainable iterative network architecture for accelerated radial multi-coil 2D cine MR image reconstruction *Medical Physics* **48** 2412–25
[29] Yang Y, Sun J, Li H and Xu Z 2016 Deep ADMM-net for compressive sensing MRI *Proc. XXX International Conference on Neural Information Processing Systems* 10–8
[30] Antholzer S and Haltmeier M 2020 Discretization of learned NETT regularization for solving inverse problems arXiv:2011.03627
[31] Li H, Schwab J, Antholzer S and Haltmeier M 2020 NETT: solving inverse problems with deep neural networks *Inverse Probl.* **36** 065005
[32] Lunz S, Öktem O and Schönlieb C-B 2018 Adversarial regularizers in inverse problems *In Advances in Neural Information Processing Systems* **31** 8507–16
[33] Mukherjee S, Dittmer S, Shumaylov Z, Lunz S, Öktem O and Schönlieb C-B 2020 Learned convex regularizers for inverse problems arXiv:2008.02839
[34] Daubechies I, Defrise M and De Mol C 2004 An iterative thresholding algorithm for linear inverse problems with a sparsity constraint *Comm. Pure Appl. Math.* **57** 1413–57
[35] Grasmair M, Haltmeier M and Scherzer O 2008 Sparse regularization with l^q penalty term *Inverse Probl.* **24** 055020
[36] Fornasier M, Naumova V and Pereverzyev S V 2014 Parameter choice strategies for multipenalty regularization *SIAM J. Numer. Anal.* **52** 1770–94
[37] Obmann D, Schwab J and Haltmeier M 2021 Deep synthesis regularization of inverse problems *Inverse Probl.* **37** 015005
[38] Han Y and Ye J C 2018 Framing U-Net via deep convolutional framelets: application to sparse-view CT *IEEE Trans. Med. Imag.* **37** 1418–29
[39] Obmann D, Nguyen L, Schwab J and Haltmeier M 2020 Sparse aNETT for solving inverse problems with deep learning *2020 IEEE XVII International Symposium on Biomedical Imaging Workshops (ISBI Workshops)* 1–4
[40] Pöschl C 2008 Tikhonov regularization with general residual term *PhD Thesis* University of Innsbruck

[41] Boyd S, Parikh N and Chu E 2011 *Distributed Optimization and Statistical Learning via The Alternating Direction Method of Multipliers* (Boston: Now Publishers Inc)
[42] Gabay D and Mercier B 1976 A dual algorithm for the solution of nonlinear variational problems via finite element approximation *Comput. Math. Appl.* **2** 17–40
[43] Glowinski R and Marroco A 1975 Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualite, d'une classe de problemes de dirichlet non lineares *RAIRO Anal. Numer.* **9** 41–76
[44] Kunisch K and Pock T 2013 A bilevel optimization approach for parameter learning in variational models *SIAM J. Imag. Sci.* **6** 938–83
[45] Wang Y, Yin W and Zeng J 2019 Global convergence of ADMM in nonconvex nonsmooth optimization *J. Sci. Comput.* **78** 29–63
[46] Adler J, Kohr H and Öktem O 2017 ODL-a python framework for rapid prototyping in inverse problems *Royal Institute of Technology* (https://github.com/odlgroup/odl)
[47] McCollough C 2016 TU-FG-207A-04: overview of the low dose CT grand challenge *Med. Phys.* **43** 3759–60
[48] Adler J and Öktem O 2018 Learned primal-dual reconstruction *IEEE Trans. Med. Imag.* **37** 1322–32
[49] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization arXiv:1412.6980